PB84-124841

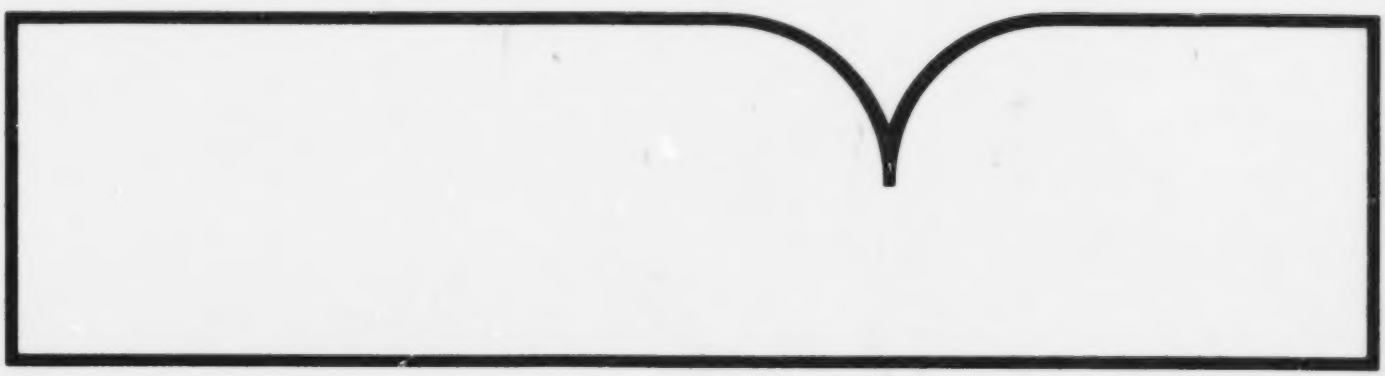
Guidelines for Conducting Single Laboratory Evaluations of Biological Methods

Bioassay Systems Corp., Woburn, MA

Prepared for

Environmental Monitoring Systems Lab.
Las Vegas, NV

Nov 83



U.S. Department of Commerce
National Technical Information Service
NTIS

PB84-124841

EPA-600/4-83-056
November 1983

GUIDELINES FOR CONDUCTING SINGLE LABORATORY EVALUATIONS OF BIOLOGICAL METHODS

by

William D. McKenzie and Theodore A. Olsson III
Bioassay Systems Corporation
Woburn, Massachusetts 01801

Contract Number 68-02-3173

Technical Project Officer
William W. Sutton
Exposure Assessment Research Division
Environmental Monitoring Systems Laboratory
Las Vegas, Nevada 89114

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U. S. ENVIRONMENTAL PROTECTION AGENCY
LAS VEGAS, NEVADA 89114

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-600/4-83-056		PB84 124841

4. TITLE AND SUBTITLE	5. REPORT DATE
Guidelines for Conducting Single Laboratory Evaluations of Biological Methods	November 1983
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) William D. McKenzie and Theodore A. Olsson III	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bioassay Systems Corporation 225 Wildwood Avenue Woburn, MA 01801	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO. 68-02-3173
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency--Las Vegas, NV Office of Research and Development Environmental Monitoring Systems Laboratory Las Vegas, Nevada 89114	13. TYPE OF REPORT AND PERIOD COVERED Final PR/PS
	14. SPONSORING AGENCY CODE EPA/600/07

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The single laboratory test is used to establish the data quality that can be achieved within a single laboratory. It provides a basis for deciding whether or not a given method merits collaborative testing and it more clearly defines a method's potential for inclusion as part of an operational monitoring network. This summary provides a brief description of the suggested procedures for single laboratory testing.

Phases of the single laboratory test include identification of procedural variables that must be carefully controlled (ruggedness testing), evaluation of method sensitivity, identification of the limits of reliable measurement, evaluation of systematic error (bias), and identification of method precision and accuracy. The chemical composition of all sample material must be verified during the single laboratory test. Sample material should have a concentration range, in the same sample matrix, that would be encountered if the method was being routinely used for its intended purpose. Some phases of the test should make use of certified reference materials.

The resulting single laboratory test data and the revised (ruggedness tested) method protocol will ultimately be used as part of the basis for deciding whether or not to proceed with collaborative testing.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Release to public		75
	20. SECURITY CLASS *(This page)*	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

The information in this document has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement #68-02-3173 to PEDCo Environmental. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

TABLE OF CONTENTS

	Page No.
LIST OF TABLES	v
LIST OF FIGURES	vi
EXECUTIVE SUMMARY	vii
INTRODUCTION	xvi
1. IDENTIFICATION OF PROCEDURAL VARIABLES WHICH MUST BE CONTROLLED	1-1
2. METHOD PRECISION	2-1
2.1 Definitions	2-1
2.2 Techniques for the Determination of Method Precision	2-1
2.3 Tabulation, Analysis, and Interpretation of Data	2-2
2.4 Outliers (also see Appendix I)	2-2
2.5 Reporting of Method Precision	2-3
3. METHOD ACCURACY	3-1
3.1 Definitions	3-1
3.2 Reference Materials	3-2
3.3 Sources of Reference Materials	3-2
3.4 Statistical Test to Determine Accuracy When Using Reference Materials	3-4
3.5 Data Evaluation of Reference Materials Over Time	3-7
4. SYSTEMATIC ERROR	4-1
4.1 Systematic Error (bias)	4-1
4.2 Identification of Systematic Error	4-2
4.3 Reporting Systematic Error	4-3
5. METHOD SENSITIVITY	5-1
5.1 Definitions and Procedures	5-1
5.2 Reporting Method Sensitivity	5-2

TABLE OF CONTENTS (Continued)

	Page No.
6. LIMITS OF RELIABLE MEASUREMENT	6-1
6.1 Estimating Limits of Reliable Measurement	6-1
6.2 Example of Reliable Measurement Estimate	6-1
6.3 Reporting the Limits of Reliable Measurement	6-2
7. DATA REQUIRED FOR A COLLABORATIVE TEST DECISION	7-1
Appendix I OUTLINERS - REJECTION OF SUSPECTED DATA	A-I-1
Appendix II ACCEPTANCE CRITERIA FOR ADEQUATE METHOD PRECISION	A-II-1
Appendix III FALSE POSITIVE AND FALSE NEGATIVE VALUES	A-III-1
Glossary	A-III-2
Notation	A-III-4
References	A-III-5

LIST OF TABLES

Table No.		Page
1-1	EXPERIMENTAL DESIGN FOR SEVEN VARIABLE RUGGEDNESS TEST	1-2
1-2	EXPERIMENTAL DESIGN FOR FOUR VARIABLE RUGGEDNESS TEST	1-4
3-1	STUDENT'S DISTRIBUTION WITH k DEGREES OF FREEDOM	3-6
Appendix I		
A-1	HYPOTHETICAL RESULTS OF BIOLOGICAL MONITORING ASSAY CONDUCTED ON TEN DIFFERENT DAYS USED TO ESTIMATE THE PRECISION OF THE ASSAY	A-I-2
A-2	CRITERIA FOR TESTING FOR EXTREME OBSERVATION (NO PAST DATA)	A-I-5
A-3	VALUES OF EXTREME DEVIATION $\frac{X_m - \bar{X}}{S_e}$ NOT REJECTED AS AN OUTLIER	A-I-7
A-4	USE OF NORMAL DISTRIBUTION	A-I-9

LIST OF FIGURES

Figure		Page
E1	Suggested flow chart for single laboratory testing	vii
E2	Example of how the evaluation analyses can provide data for multiple portions of the single laboratory test	viii
E3	Examples of different reference material concentrations used in sensitivity testing	xii
3-1	Control chart for reference sample	3-3
5-1	Example of different reference material concentrations used in sensitivity testing	5-1

EXECUTIVE SUMMARY

This guideline document is intended to provide specific guidance for conducting a single laboratory test. Single laboratory testing is used to establish the data quality that can be achieved within a single laboratory. It provides a basis for deciding whether or not a given method merits collaborative testing and it more clearly defines a method's potential for inclusion as part of an operational monitoring network.

The single laboratory test includes identification of procedural variables that must be carefully controlled (ruggedness testing), evaluation of method sensitivity, identification of the limits of reliable measurement, evaluation of systematic error (bias), and identification of method precision and accuracy.

A complete protocol, for the method being evaluated, must be received by the evaluating laboratory prior to test initiation. All method requirements and procedural instructions should be clearly presented. It is obviously important that these written instructions be technically correct, complete, and as unambiguous as possible. However, the laboratory conducting the single laboratory test is not responsible for the actual method itself. The laboratory must strictly follow the method procedures and method requirements (experimental conditions, reagents, laboratory equipment, storage of samples, maintenance of experimental organisms, blanks, standards, replicates, etc.) as they are written in the protocol.

For the purposes of this document, biological methods will include procedures used to chemically or biochemically analyze biological tissues and fluids as well as the various biological analyses for toxicity, mutagenecity, etc. Single laboratory test objectives will be somewhat different depending on the method that is being evaluated. For example, to determine a method's capability for accuracy (and for systematic error), the testing laboratory must use a reference sample material which has a known response (true value) available for that method. The chemical or biochemical analysis of tissues would use a reference material with known compound concentrations or perhaps with certified enzymatic activity levels. However, when the technique being evaluated includes a toxicity test such as an algal assay (population stimulation or inhibition) there is frequently no "true value" or "true response" and hence the method's single laboratory capability for accuracy can not really be determined. Under these conditions, an average test response should be acquired by conducting successive analyses on the same concentration of the same reference material. Because of such differences, the specific requirements for each single laboratory test must always be confirmed with the sponsor before single laboratory testing begins.

Ten successive analyses (i,e., acquiring ten valid responses by following the method protocol) have been suggested for several phases of the single laboratory test. The method precision determination, for example, requires that ten successive independent analyses be conducted on the same sample material. Multistage calculations to determine the required number of analyses might be conducted during the single laboratory test as more information becomes available on the expected variance. However, ten analyses will allow the test laboratory to estimate the standard deviation to

within 45 percent of it true value (at a 95 percent confidence interval). The single laboratory test cost will rapidly increase as more of these successive analyses are required since each additional value must represent a valid test response and therefore will include whatever is required (blanks, replicates, etc.) in the original method protocol to insure a valid test response.

The following summary provides a brief description of the instructions for single laboratory testing. Before the single laboratory test begins, the evaluating laboratory must review the protocol and make notations where ambiguous statements are made or where more detail is needed. These questions (or recommendations for improvement) must be resolved with the sponsor, and if necessary, a second protocol prepared, before the actual laboratory test begins. The single laboratory test should begin with ruggedness testing but it is important that the laboratory performing the single laboratory test plan all the assays required in advance in order to prevent duplication of effort. Some of the evaluations may be performed simultaneously, depending on the nature of the test methods being investigated.

Figure E-1 summarizes the method characteristics described in the following pages in the form of a flow chart which may be followed during the single laboratory test. An example of how the evaluation analyses can provide data for multiple portions of the single laboratory test is illustrated in Figure E-2. This figure illustrates how the different analyses can provide data for more than one portion of the single laboratory test. This figure assumes that a single test material is being used i.e., different concentrations of the same reference material. Many reference materials are available and the type of sample materials used should always be discussed with the sponsor prior to the initiation of a single laboratory test.

Ruggedness Testing

The single laboratory test should first identify any procedural variables that must be carefully controlled (ruggedness testing). If the given method is "rugged" it will not be susceptible to the inevitable, modest departures in routine and the results obtained will not be altered by these minor variations. If the results are altered by small procedural variations, it is important to emphasize in the protocol that a specific step must be strictly followed or, in some cases, to indicate the limits of allowable variability. For example, the ruggedness test results might indicate that a certain (protocol directed) temperature requirements of 20°C was critical and that slight variations in this temperature (at a given phase of the test) would produce an altered test result. The method protocol will then need to be revised to emphasize that the stated temperature requirement must be strictly followed and perhaps, to provide more detail on any quality control steps associated with temperature determinations. Depending on the sponsor's requirements, additional tests might be conducted to determine the specific narrow range of temperatures that would be allowable i.e., a protocol revision noting that the temperature can not vary except by a stated amount. The minor departures in routine selected for ruggedness testing do not always need to be variations in the protocol. Unjustified latitude is sometimes given in a protocol simply because of insufficient detail.

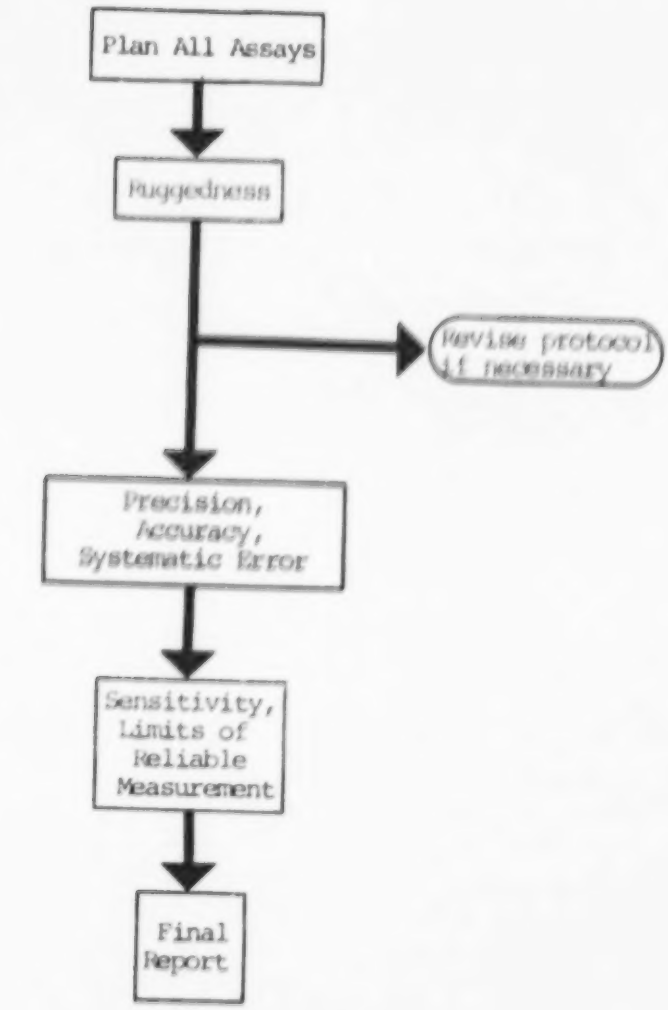
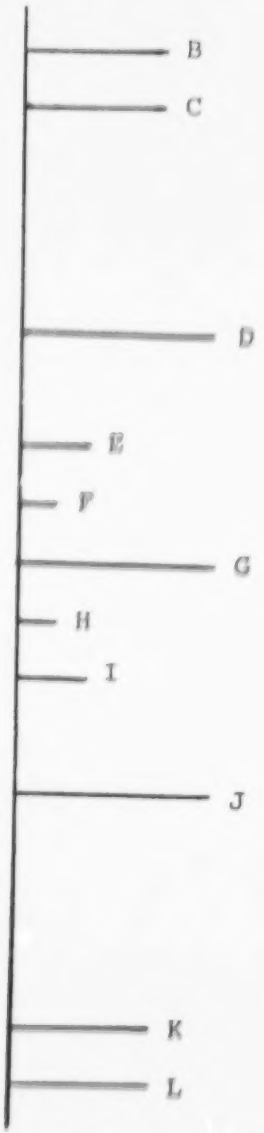


Figure E-1. Suggested flow chart for single laboratory testing



Precision - G

Method Sensitivity - D,E,F,H,I,J

Limits of Reliable Measurement

1. precision B,L/compared with data from G
2. sensitivity B,C,L,K/compared with data from E,F,H,I,

Accuracy - G (10 additional valid responses should be acquired if no true response is available i.e., total of 20).

Systematic Error-D,G,J (these might be used if a true response is available, if no true response, then no data will be acquired for systematic error.

Figure E-2. Example of how the evaluation analyses can provide data for multiple portions of the single laboratory test.

The ruggedness test would probably use concentration G but should vary the experimental procedure for the different analyses. The above chart assumes a ruggedness tested/revised protocol is being used. Each letter represents 10 successive independent analyses.

The suggested approach does not seek to study each separate variable in an individual sequential fashion, but rather it provides for the introduction of several changes (protocol variations) simultaneously. At least seven variables should be selected which will require that the test laboratory conduct eight separate analyses. However, one of these variables could be a meaningless variable thus providing a modified control for the ruggedness test itself. (The experimental design for a seven variable test is presented in the following section on ruggedness testing.) Basically, the differences between the "protocol directed" result and the "protocol altered" result for each variable are compared. If one or two variables are having an effect on the test result, their respective differences (directed vs altered) will be substantially larger than the group of differences associated with the other variables. Most of the modest procedural alterations should have little effect on the test result since the variations should only be of a magnitude that could be made by a qualified laboratory following the written method protocol.

The number of variables to select should be discussed with the sponsor prior to the test. However, reference tables (for designing ruggedness tests) are available when certain numbers of procedural variables are selected (see Section 1). Having completed the ruggedness evaluation, the subsequent phases of the single laboratory test (precision, method sensitivity, etc.) can be conducted using the revised method procedure.

Method Precision

The only requirement for the precision test is to conduct 10 separate analyses on the same sample (preferably using a reference material). Each separate analysis must represent a valid test response as required by the particular method protocol. It is also recommended that the separate precision analyses be conducted on alternate days (i.e., an interval of at least one day between the completion of one anaylsis and the start of the next). The resulting data can be expressed either as a standard deviation, a standard error, or as a coefficient of variation.

Method Accuracy

To determine a method's single laboratory capability for accuracy (and for systematic error), the testing laboratory must have both a standard reference material and a known method response (true response) to this reference material. When a method calls for analysis of biological tissue or biological fluid, there will usually be a standard reference material available to the testing laboratory e.g., samples with certified compound concentrations or perhaps with certified enzymatic activity levels. In these instances, the method's single laboratory capability for accuracy can be assessed by determining the differences between the observed single laboratory result, using the reference sample, and the known true value. Ten separate determinations (10 valid responses) should be conducted using a single concentration of the reference sample material. Each of the ten determinations must represent a valid test response as directed in the

particular method protocol. The method protocol presents whatever requirements are necessary for replicates, blank samples, etc., in order to provide a valid response. The Student t-Test would be used to determine the significance of the difference between the observed single laboratory test result and the known true value.

If the method being single laboratory tested is a bioassay test such as an algal assay test (population stimulation or inhibition), there will usually be no "true response" available for a reference material and hence the method's single laboratory capability for accuracy (or for systematic error) can not be determined. Under these conditions, the testing laboratory should first select a reference material, with sponsor's approval, and then determine an average test response for a single concentration of the reference sample. When determining the average test response, it is recommended that 20 independent determinations (valid test responses as indicated by method protocol) be conducted. While the literature data base may provide valuable background information for a method's average response to various sample types, the single laboratory testing group should still conduct these determinations using the ruggedness tested/revised protocol.

Systematic Error

A method's capability for (minimizing) systematic error can be considered as a part of the method's capability for accuracy. If a true response (known value) is not available for a reference material, the single laboratory test will not be able to acquire data on the method's systematic error. The testing laboratory should notify the sponsor that, under these conditions, no data will be provided for this phase of the single laboratory test. Comparison of the test data with the results of a reference method can be used in certain situations, but for the purpose of this document, the use of such reference methods are not considered as part of single laboratory testing. It should also be remembered that it is the bias of the method, not the bias of the laboratory that is being addressed. Single laboratory testing does not address laboratory bias even though it will obviously affect test results.

In the case of a bioanalytical method for which a standard reference material is available, the testing laboratory should estimate a method's systematic error by using various dilutions of the reference sample. A single concentration of the standard reference material would be aliquoted into at least three equal samples. Two of these aliquots would then be diluted to different total volumes, thus creating three different sample sizes. Ten respective independent analyses would then be conducted on each of the three sample groups. Results from each group (results from two groups would first be corrected by the respective dilution factor) would then be compared (Student's t-Test) with the true value for the reference material. The consistently present systematic error should be noted in each of these three test groups. Additional concentrations or dilutions as well as additional independent determinations would probably be beneficial for the single laboratory assessment of systematic error (depending on the sponsor's requirements).

Figure E-3 illustrates the selection of concentrations for sensitivity testing. Assume concentration A had been selected as the concentration of sample material used previously in the method precision test. For the sensitivity test, the laboratory would select one concentration greater than (C), and one concentration less than (B) the concentration used during the precision test. These concentrations should probably be equally distant from the precision test concentration (A). The laboratory should conduct 10 independent analyses for each new concentration (i.e., 10 separate valid responses acquired by following the method protocol). Assuming that the procedure can distinguish between A and C, and between A and B, then the test laboratory will reduce the concentration interval by one-half (to C' and B'). If the method is still capable of distinguishing between A and C', and between A and B', then the test laboratory will again reduce the sample concentration interval by one-half (to C" and B"). The 10 independent analyses on these last respective concentrations (C" and B") will complete the sensitivity test even if the method can distinguish between A and C", and between A and B". The sponsor can indicate if additional information on the method's sensitivity is required and, if so, direct the test laboratory to continue this process or to repeat the process using a different reference material.

A relatively poor method capability for sensitivity does not necessarily imply limited method usefulness or that the method would be an unlikely candidate for collaborative testing. The intended purpose of the respective method must always be considered when reviewing single laboratory test data.

Limits of Reliable Measurement

In determining a method's limits of reliable measurement, the single laboratory test data may simply verify that the method capabilities for sensitivity, precision, and accuracy (if applicable) do not deteriorate at the upper and lower extremes of the detection range. The same reference material should be used for these tests as was originally used during the method precision and method sensitivity evaluations.

The single laboratory test is not required to establish an upper and lower detection limit. It is assumed that a sufficient literature data base for the method exists to grossly estimate an upper and lower limit of detection using the particular sample material. The evaluating laboratory should initially select two concentrations of the sample material. One of these concentrations will be near the upper extreme of the method's detection range and the other concentration will be near the lower extreme of the detection range. Ten analyses would be conducted on each concentration to provide precision data (expressed as a coefficient of variation). The coefficient of variation will frequently show a dramatic increase at the extreme limits of detection and, therefore, the precision data will provide a distinct indication of the limits of reliable measurement. Two additional concentrations, one at each extreme of the estimated response range should then be selected in order to conduct the sensitivity determinations. Thus, two sample concentrations at each extreme of the estimated response range can

Method Sensitivity

For purposes of a single laboratory test, a method's sensitivity is defined as the method's capability to detect (or distinguish between) small changes in sample concentration (i.e., concentration of analyte). A chemically characterized reference material should be used as sample material during sensitivity testing.

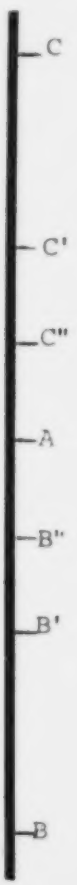


Figure E-3. Example of different reference material concentrations used in sensitivity testing.

be compared (in terms of precision and sensitivity) with the previously acquired data. If the method is a bioanalytical technique with an available "true value" for the reference material, accuracy data would also be acquired.

When using this evaluation plan, a true limit of reliable measurement may not actually be established. However, even under these conditions, data would still be available to indicate that the technique was (or was not) capable of providing reliable measurements at the extreme concentrations selected. Additonal sample concentrations, or additional test substances, can also be selected based on the sponsor's needs.

INTRODUCTION

In order for a method to be selected for use in an operational monitoring network (e.g., biological tests used to implement RCRA requirements as well as most monitoring techniques as per EPA mandatory quality assurance requirements), it must be characterized as to its capability for ruggedness, precision, sensitivity, limits of reliable measurement, systematic error, and accuracy. The method's reproducibility must also be evaluated after the method is used by various qualified laboratories (collaborative testing).

Collaborative testing is ultimately necessary before a method can be reliably used. However, collaborative testing is expensive and very time consuming. It is important that single laboratory test data be available as a basis for selecting which method will be collaboratively tested. The single laboratory test includes identification of procedural variables that must be carefully controlled (ruggedness testing), evaluation of method sensitivity, identification of the limits of reliable measurement, evaluation of systematic error (bias), and identification of method precision and accuracy. For the purposes of this document, biological methods will include procedures used to chemically and biochemically analyze biological tissues and fluids as well as the various biological tests for toxicity, mutagenecity, etc. Single laboratory test objectives may be somewhat different depending on the method that is being evaluated.

A complete protocol, for the method being evaluated, must be received by the evaluating laboratory prior to the initiation of the test. The laboratory must strictly follow the method procedures and method requirements (experimental conditions, reagents, laboratory equipment, storage of samples, maintenance of experimental organisms, blanks, standards, replicates, etc.) as they are written in the protocol.

After the testing laboratory has reviewed the method protocol and has resolved any problem areas with the sponsor, the ruggedness testing phase should be conducted first. It is assumed that sufficient data, a previously accumulated data base, is available to estimate a gross maximum and minimum response range as well as to provide some information about the type of test response that will probably result when using the planned sample material. During ruggedness testing, the protocol will probably be improved by revision and clarification and the subsequent test phases (method sensitivity, limits of reliable measurement, etc.) will be conducted using the revised procedure.

The single laboratory test cannot evaluate a method's scope of applicability, i.e., the array of sample types or environmental situations that the method would be able to provide useful data for. However, the single laboratory test should include various test samples having a concentration range, in the same sample matrix, that would be encountered if the method was being routinely used for its intended purpose. Chemical composition of the sample material should be verified during the single laboratory test.

In order to make efficient use of both time and available funding, the test laboratory, and the sponsor, should carefully plan all of the anticipated assays in advance. Some of the test phases can be performed simultaneously, and some of the resulting data can provide information for more than one test phase.

The U.S. Environmental Protection Agency contracted PEDCo Environmental and Bioassay Systems Corporation to prepare this document under U.S. EPA Contract No. 68-02-3173, PEDCo Subcontract No. 3450-73. The document was prepared according to the U.S. EPA requirements for single laboratory testing as specified by Dr. William W. Sutton, U.S. EPA Environmental Monitoring Systems Laboratory/Environmental Assessment Division, Las Vegas, Nevada. Bioassay Systems gratefully acknowledges the guidance and cooperation provided by Dr. Sutton in the preparaton of this document.

1. IDENTIFICATION OF PROCEDURAL VARIABLES WHICH MUST BE CONTROLLED

A complete protocol, for the method being evaluated, must have been received by the evaluating laboratory prior to test initiation. All requirements and procedural instructions must be clearly presented. It is obviously important that these written instructions be technically correct, complete, and as unambiguous as possible. However, the laboratory conducting the single laboratory test is not responsible for the actual method itself. The laboratory must strictly follow the method procedures and method requirements (experimental conditions, reagents, laboratory equipment, storage of samples, maintenance of experimental organisms, blanks standards, replicates, etc.) as they are written in the protocol.

Before the single laboratory test begins, the evaluating laboratory must review the protocol and make notations where ambiguous statements are given, or where more detail is needed. These questions (or recommendations for improvement) must be resolved with the sponsor, and a revised protocol prepared, before the actual single laboratory test begins.

Although the approved test protocol will be used, procedural variability which affects "ruggedness" may still exist. If the given method is "rugged", it will not be susceptible to the inevitable, modest departures in routine and the results obtained should not be altered by minor departures. For biological tests these variables (or departures) often occur in the area of environmental control. For example, slight changes in temperature and light period or intensity. If the results are altered by small variations, it is important to identify the range of variation which will produce unaltered results. The minor departures in routine selected for ruggedness testing do not always need to be variations in the protocol. Unjustified latitude is sometimes given in a protocol simply because of insufficent detail.

The intentionally introduced departures must be small and of a magnitude similar to the modifications expected when different analysts in various qualified laboratories use the method over different runs or days. If the results are altered after a modest departure for the procedure, the protocol must be revised to emphasize that the method user must strictly follow the stated method protocol.

A single concentration of one test material can be used for the ruggedness test. The suggested approach does not seek to study each separate variable in an individual sequential fashion but rather it provides for the introduction of several changes (protocol variations) simultaneously. At least seven variables should be selected which will require that the test laboratory conduct eight separate analyses. However, one of these variables could be a meaningless variable thus providing a modified control for the ruggedness test itself. Basically, the difference between the "protocol directed" result and the "protocol altered" result for each variable are compared. If one or two variables are having an effect on the test result, their respective differences (directed vs. altered) will be substantially larger than the group of differences associated with the other variables. Most of the modest procedural alterations should have little effect on the test result since the variations should only be a magnitude that could be made by a qualified laboratory following the written method protocol.

Table 1-1 summarizes a ruggedness test using seven variables which will therefore require eight separate analyses. The varied condition is to be either slightly above or slightly below the "protocol directed" condition. The "protocol directed" conditions are designated as A to G, and the varied conditions are designated as a through g. The evaluation is concerned with identifying respective variations in the final test result due to the specific procedural differences, i.e., A-a, B-b, C-c, D-d, E-e, F-f, and G-g. Each of the eight trials consists of a single analysis conducted using eight respective aliquots of a single test material. The final results are indicated as s, t, u, v, w, x, y and z.

TABLE 1-1 EXPERIMENTAL DESIGN FOR A SEVEN VARIABLE RUGGEDNESS TEST

Number	Factor level	Average result
	A B C D E F G	
1	A B c D e f g	s
2	A b C d E f g	t
3	A b c d e F G	u
4	a B C d e F g	v
5	a B c d E f G	w
6	a b C D e f G	x
7	a b c D E F g	y
8		z

To determine whether changing factor A to a has had an effect on the test result, the average $\underline{A} = (s + t + u + v)/4$ will be compared with the average of $\underline{a} = (w + x + y + z)/4$. Since each of the two groups of four determinations contains the other six variables, twice at the upper case level and twice at the lower case level, the effects of these variables (if present), cancel out, leaving only the effect of changing variable A to a. The other variable can also be evaluated by examining the following averages:

$$B = \frac{(s + t + w + x)}{4} \qquad b = \frac{(u + v + y + z)}{4}$$

$$C = \frac{(s + u + w + y)}{4} \qquad c = \frac{(t + v + x + z)}{4}$$

$$D = \frac{(s + t + y + z)}{4} \qquad d = \frac{(u + v + w + x)}{4}$$

$$E = \frac{(s + u + x + z)}{4} \qquad e = \frac{(t + v + w + y)}{4}$$

$$F = \frac{(s + v + w + z)}{4} \qquad f = \frac{(t + u + x + y)}{4}$$

$$G = \frac{(s + v + x + y)}{4} \qquad g = \frac{(t + u + w + z)}{4}$$

After tabulating the above averages, the difference between each respective variable is computed, e.g.,

$$A - a = \frac{(s + t + u + v)}{4} - \frac{(w + x + y + z)}{4}$$

Examination of these respective differences enables the evaluating laboratory to determine which variables are effecting the test result. As stated above, most of the method's procedural alterations should have little or no effect on the result. Considerable information can be gained by merely comparing these differences (A-a, B-b, C-c, D-d, E-e, F-f, and G-g).

The evaluating laboratory may wish to conduct multiple tests (repeat analyses) for each variable combination (e.g., 10 successive independent analyses for each of the 8 combinations) depending on the sponsor's requirements. Under these conditions, if any of the respective differences between averages are greater than two times the standard deviation for each variable (experiment), the testing laboratory will have another indication that the particular variable is effecting the test result.

After discussing the results with the sponsor, additional studies might be planned to define the limits of acceptable variation for a particular critical test procedure.

Experimental designs using other numbers of variables are also available. However eight experimental groups are still used for an evaluation of four factors (Table 1-2). As before, the capital letters A, B, C, and D denote the "protocol directed" instructions which, if altered, may cause a change in the test result. Their "altered condition" is denoted by the respective lower case letters a, b, c, and d.

TABLE 1-2. EXPERIMENTAL DESIGN FOR A FOUR VARIABLE RUGGEDNESS TEST

Experiment Number	Combination of Conditions	Results
1	A B C D	s
2	A B c D	t
3	A b C d	u
4	A b c d	v
5	a B C d	w
6	a B c d	x
7	a b C D	y
8	a b c D	z

The following theoretical example is provided for illustrative purposes only. The test conditions for this evaluation of ruggedness were selected as slight variations in the test protocol for a biological test method measuring LC_{50} values:

Condition	Normal Condition	Value of Condition
Temperature	A = 21°C	a = 25°C
Period of Light Exposure	B = 10 hrs.	b = 8 hrs.
Salinity of Test Media	C = 3.0%	c = 1.0%
Dissolved oxygen content	D = 60%	d = 40%

When the above conditions were combined as specified in Table 1-2, the following averages of ten independent results were observed:

Experiment Number	Letter Code (see Table 1-2)	Average LC 50 Result (n = 10)	Standard Deviation (S) For n = 10 measurements
1	s	25%	2.6
2	t	15%	2.4
3	u	23%	2.5
4	v	16%	2.6
5	w	24%	2.8
6	x	16%	2.2
7	y	26%	3.0
8	z	14%	2.3

The following averages of the average results for each experiment were calculated.

$$A = \frac{(s + t + u + v)}{4} = \frac{(25 + 15 + 23 + 16)}{4} = 19.75$$

$$a = \frac{(w + x + y + z)}{4} = \frac{(24 + 16 + 26 + 14)}{4} = 20.00$$

$$B = \frac{(s + t + w + x)}{4} = \frac{(25 + 15 + 24 + 16)}{4} = 20.00$$

$$b = \frac{(u + v + y + z)}{4} = \frac{(23 + 16 + 26 + 14)}{4} = 19.75$$

$$C = \frac{(s + u + w + y)}{4} = \frac{(25 + 23 + 24 + 26)}{4} = 24.50$$

$$c = \frac{(t + v + x + z)}{4} = \frac{(15 + 16 + 16 + 14)}{4} = 15.25$$

$$D = \frac{(s + t + y + z)}{4} = \frac{(25 + 15 + 26 + 14)}{4} = 20.00$$

$$d = \frac{(u + v + w + x)}{4} = \frac{(23 + 16 + 24 + 16)}{4} = 19.75$$

The difference between the averages were calculated as follows:

$$A - a = 19.75 - 20.00 = -0.25$$

$$B - b = 20.00 - 19.75 = +0.25$$

$$C - c = 24.50 - 15.25 = +9.25$$

$$D - d = 20.00 - 19.75 = +0.25$$

Inspection of the above differences confirm the drastic effect of changing the salinity. The large + difference indicates much higher toxicity at 3.0% salinity versus 1.0%.

The significance of this difference can be further demonstrated by comparing the difference to 2 x the standard deviation (S) of the mean value for each experiment. If the difference is greater than 2 S then the difference is concluded to be significant.

Experiment Number	Calculation	Conclusion*
1	2 x (2.6) = 3.7	significant
2	2 x (2.4) = 3.4	significant
3	2 x (2.5) = 3.5	significant
4	2 x (2.6) = 3.7	significant
5	2 x (2.8) = 4.0	significant
6	2 x (2.2) = 3.1	significant
7	2 x (3.0) = 4.2	significant
8	2 x (2.3) = 3.3	significant

*When the difference of 9.25 (C-c) is greater than 2 X S, the difference is significant.

The obvious conclusion of the ruggedness testing is that the salinity concentration of 3.0% grossly effected the results. The results of this test indicated a significant difference between these levels of salinity, therefore the method was concluded not to be "rugged". The following statement was made in the final report concerning the minimum range of conditions which did not effect the test method for the test material analyzed.

Temperature	21 - 25°C
Light exposure	8 - 10 hours
Salinity	1.0%
Dissolved oxygen	40 - 60%

As was stated above, the ruggedness test should be conducted using procedural variables that might actually occur when a qualified laboratory is following the written protocol. A single concentration of sample material can be used throughout the ruggedness test but all test sample material should be approved by the sponsor. At least seven separate variables should be selected. Based on the sponsor's requirements, additional studies can be conducted to define the limits of acceptable variations for particular critical test procedures.

2. METHOD PRECISION

2.1 Definitions

Precision may be defined as a measure of the repeatability and/or reproducibility of multiple measurements performed on the same sample. There are a number of statistical procedures designed to estimate precision, all of which provide information with regard to the confidence one may have in the repeatibility of experimental results. The dispersion around the mean, obtained when a sample is analyzed by several laboratories using the same method, is considered a measure of total between-laboratory precision, and is frequently defined as reproducibility. Reproducibility is, therefore, an overall measure of variability, including both within-laboratory and between-laboratory components. Repeatability is the degree of variability of replicate samples analyzed during the same time period as well as on future analysis dates. It is the measure of within laboratory random, indeterminate error. Results that include random and normally distributed errors are obtained when an analyst carries out a number of replicate determinations on the same sample with the same procedure, test system, reagents, conditions, etc. Since the results of replicate determinations are considered to be a random sample from a normal population of results obtained in the same way, the estimated standard deviation (S) or, in some cases, the coefficient of variation of this distribution (CV) is used to express the precision of an experimental method.

2.2 Techniques for the Determination of Method Precision

Single laboratory precision testing of biological monitoring methods will focus on quantitating the distribution of individual measurements, within a defined data set, around their average or mean value. A typical study to evaluate method precision would include testing a representative sample or appropriate reference material according to the approved method protocol. The test would be performed with a minimum of 10 successive measurements made on the same test material and would include the number of replicate analyses blanks, standards, etc. required by the method protocol. The data obtained should be evaluated by calculating the standard deviation about the mean value and the coefficient of variation (or relative standard deviation). These measurements of variation will serve as an index of method precision for single laboratory testing.

2.3 Tabulation, Analysis and Interpretation of Data

An estimate of the standard deviation, based upon the results obtained from a minimum of 10 successive independent analyses on the same test material may be calculated using Bessel's Correction (Neville and Kennedy, 1964), according to the following general formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad \text{Eq. 2-1}$$

Where: S = Standard deviation
n = Number of results
X_i = Individual results
$\bar{X}$ = Average result (mean)

In this equation, the standard deviation (S) is an estimate of the population standard deviation. This value is a measure of the dispersion of a series of results around their average (X), expressed as the square root of the quantity obtained by summing the squares of the deviations from the average of the results and dividing by the number of observations minus one.

Each of the above mentioned 10 independent analyses will be performed with a minimum 1 day interval between the last analysis and the next using the number of replicates, concentrations, blanks and controls required by the method protocol.

The coefficient of variation will also be used to express precision of the method in relative terms, as percentage, instead of the absolute units of standard deviation. The coefficient of variation is expressed as follows:

$$CV = \frac{S}{\bar{X}} \times 100 \qquad \text{Eq. 2-2}$$

Where: CV = Coefficient of variation
S = Standard deviation
$\bar{X}$ = Mean

2.4 Outliers

A valid estimation of precision will include the elimination of outliers. Outliers are observed results which lie outside of the statistical frequency distribution in a series of analytical values, i.e., those values that do not follow the developing trend of other data values in the group. If an outlier is suspected, the test laboratory is directed to Appendix I, Outliers - Rejection of Suspect Data.

2.5 Reporting of Method Precision

In order to interpret the results of a method, results obtained from any type of measurement must include clear statements concerning their uncertainty. The report of precision must contain the standard deviation (S), the coefficient of variation (CV), the sample size (a minimum of 10 independent measurements), a description of the test method and the identity of the test sample. The chemical composition of the test sample should also be reported.

Reported results should be given, after rounding, to the appropriate number of significant figures consistent with the value in the calculations containing the least significant figures. It is suggested that the lack of precision should be stated to at least three figures, but not more than the significant figures in the mean.

Whenever a reliable determination of results is available, it should be reported in separate statements. Since the term "±b" in the expression a±b may be interpreted either as bounds to the inaccuracy of a, standard deviation, or standard error of the mean, it is required that the short hand "a±b" be avoided, unless accompanied by an explicit explanation of its connotation.

3. METHOD ACCURACY

To determine a method's capability for accuracy (and for systematic error), the testing laboratory should use a certified reference material. When a method calls for the chemical or biochemical analysis of biological tissue or biological fluid, there will usually be a standard reference material available to the testing laboratory, e.g., samples with certified compound concentrations or perhaps with certified enzymatic activity levels. However, when the technique being evaluated includes a bioassay test such as an algal assay (population stimulation or inhibition) there is no "true value" or "true response" and hence the method's single laboratory capability for accuracy cannot be determined. Under these conditions, an average test response should be acquired by conducting successive analyses on the same concentration of the same reference material. Because of such differences, the specific requirements for each single laboratory test must always be confirmed with the sponsor before single laboratory testing is initiated.

Accuracy data (where apprropriate) must be compiled by the single laboratory before any attempt at collecting data is warranted by a collaborative effort. The accuracy data compiled by the single evaluating laboratory can then be confirmed by obtaining consensus values of 5 or more independent laboratories (Horwitz et al, 1978). This step should only be undertaken if data concerning precision, systematic error, ruggedness, sensitivity and limits of reliable measurement have already been evaluated.

3.1 Definitions

Accuracy can be defined as the difference between the experimental measurement and its true, accepted value. It can be expressed relative to the true value as follows:

$$\frac{\left| X_{Exp} - X_T \right|}{X_T} \times 100 \qquad \text{Eq. 3-1}$$

Where:

X_{Exp} = Experimental measurement

X_T = True value

For most biological methods, the difference between the experimental and true value may be difficult, if not impossible, to measure due to the absence of true response values for a reference material. The reference materials should be certified (National Bureau of Standards, Environmental Protection Agency, etc.) and must be available for use by other laboratories.

In these cases, accuracy can not be evaluated and the single test laboratory is directed to establish a "true" value for a given method and reference material.

The "true" value will be established by performing 20 successive, independent analyses, each analysis containing the number of replicates and controls required by the test protocol to produce a valid response. This test may be performed simultaneously with those of the precision section (Section 2).

Whenever a true value is being established report the data in the form of a control chart with the 2S and 3S limits plotting the experiment mean and standard deviation of each independent analysis. Use the same model illustrated in Figure 3-1 with the exception that a true value will not be available. After plotting all 20 analytical results, note if there is a distinct increase or decrease (with time) about the total mean of the 20 results. If there is, this may be an indication of the emergence of a systematic error during testing. Note all observations in the final report.

3.2 Reference Materials

For the purposes of single laboratory testing (biological methods), the measurement of accuracy of the test method should be performed using reference materials. As stated above, reference materials will have a certified chemcial composition and be available to other test laboratories.

Suitable standard reference materials must have the following characteristics to evaluate the accuracy of a method:

1. Materials used must be well characterized.

2. Materials used must be of a known concentration at time of testing (stable under storage conditions, etc.) and should be homogeneous.

3. Test system response to the materials should be known (previously established data should be available) and highly repeatable under constant test conditions.

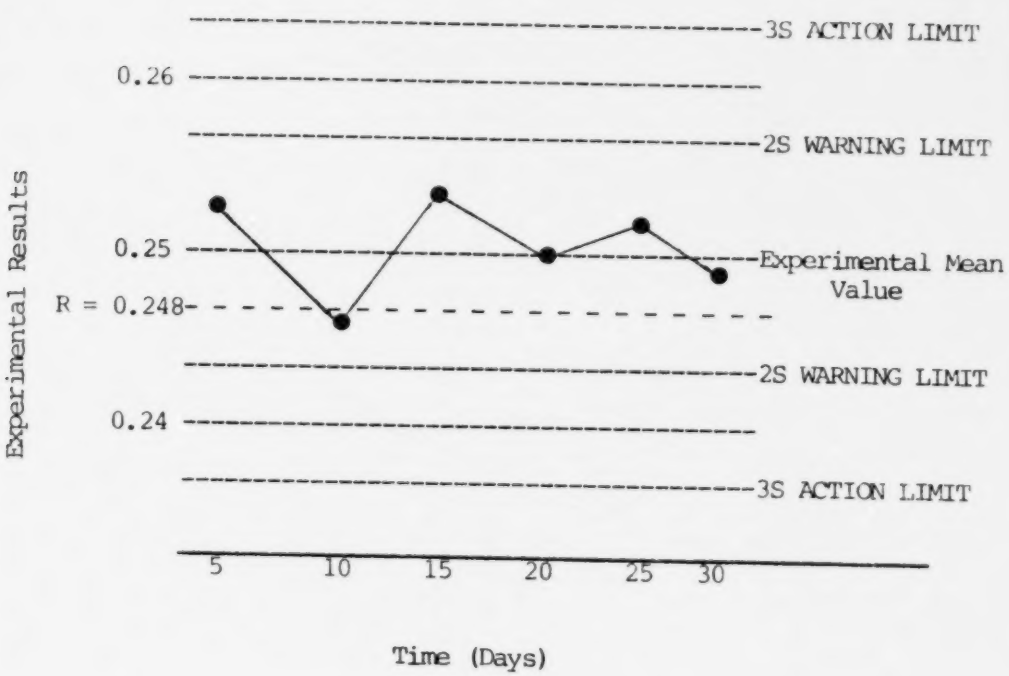
4. Materials should be easy to obtain.

Reference materials must be chemically re-analyzed after the accuracy test to confirm actual composition and identify any possible contamination or degradation.

3.3 Sources of Reference Materials

The NBS has certified a number of environmental Standard Reference Materials (SRM) (e.g., fossil fuels, water, urban dusts, gas mixtures and biological materials) which may be applicable to some biological monitoring methods (Alvarez et al, 1982). If SRM are not available for some biological monitoring methods, it may be necessary to construct a synthetic sample for evaluating method accuracy. The proportions of known or suspected hazardous component(s) to inert sample matrix components should simulate that present in the field samples to be tested in routine use of the method.

The EPA repository of hazardous materials is available from which to select candidates for evaluating method accuracy. Also, The National Bureau of Standards (NBS) may provide the appropriate reference material. For some analyses, the use of radiolabelled SRM may be desirable due to the relative ease of concentration analysis. All reference materials must be approved by the sponsor prior to testing.



R = known response value for reference material.

Figure 3-1. Example of a control chart for a hypothetical reference sample

3.4 Statistical Test to Determine Accuracy When Using Reference Materials with Known Response Values

When it is possible to utilize a reference material to estimate method accuracy, it is necessary to determine to what degree the test value differs from the known reference response value. This must be accomplished by performing a minimum of 20 successive, independent analyses on the reference material with the number of replicates and controls required by the test protocol to measure a valid response. Depending on the method being investigated, it may be possible to integrate the testing required to estimate both precision and accuracy. The single test laboratory must plan all assays in advance to reduce duplication of effort.

If the experimental values are X_1, X_2, ... X_{20} with mean $\bar{X}$, and standard deviation $S_{\bar{X}}$, then the Student's t-Test should be used to test the significance of the difference between the mean X and reference value, R. For those biological methods (i.e. bioanalytical methods) where a true value is available for a given reference material the accuracy of the method for testing that reference material will be calculated as illustrated in Eq. 3-2. The significance of the difference between experimental and true value will be evaluated using the Student's t Test.

$$t_{(\alpha,k)} = \frac{|\bar{X} - R|\sqrt{n}}{S_{\bar{X}}} \qquad \text{Eq. 3-2}$$

where:

R = reference value
$\bar{X}$ = mean experimental value
α = level of significance (confidence level)
k = n-1, degrees of freedom
n = number of samples tested
$S_{\bar{X}}$ = standard deviation

The t-test expresses the difference $(\bar{X}-R)$, as a multiple of the standard deviation of the mean, $\bar{X}$. After finding a value for t, one can then determine, using probability tables for the t-test (Table 3-1), the significance of the observed difference between the experimental and reference value.

For example: $n = 20$ replicates
$\bar{X} = 48$
$R = 52$
$S_{\bar{X}} = 3$
$k = 20-1=19$

$$\text{then: } t = \frac{|48 - 52| \times \sqrt{20}}{3}$$
$$= 5.96$$

From Table 3-1, if the calcuated t = 5.96 for 19 degrees of freedom (k=20-1) and is greater than the critical t value, $t_c = 2.09$, then the probability that the difference between $\bar{X}$ and R is significant is at least 95%. In this case, the method may be considered inaccurate. The 95% confidence interval about the experimental mean was calculated as follows:

$$\bar{X} \pm t_c \frac{S}{\sqrt{n}}$$

$$48 \pm 2.09 \frac{3}{\sqrt{20}} = 48 \pm 1.40$$

Table 3-1. PERCENTILE VALUES (t_c) FOR

STUDENT'S t DISTRIBUTION WITH k DEGREES OF FREEDOM

Degrees of freedom k	Level of Significance α (%)					
	1	2	5	10	20	Two-sided
	0.5	1	2.5	5	10	Single-sided
1	63.66	31.82	12.71	6.31	3.08	
2	9.92	6.96	4.30	2.92	1.89	
3	5.84	4.54	3.18	2.35	1.64	
4	4.60	3.75	2.78	2.13	1.53	
5	4.03	3.36	2.57	2.02	1.48	
6	3.71	3.14	2.45	1.94	1.44	
7	3.50	3.00	2.36	1.90	1.42	
8	3.36	2.90	2.31	1.86	1.40	
9	3.25	2.82	2.26	1.83	1.38	
10	3.17	2.76	2.23	1.81	1.37	
11	3.11	2.72	2.20	1.80	1.36	
12	3.06	2.68	2.18	1.78	1.36	
13	3.01	2.65	2.16	1.77	1.35	
14	2.98	2.62	2.14	1.76	1.34	
15	2.95	2.60	2.13	1.75	1.34	
16	2.92	2.58	2.12	1.75	1.34	
17	2.90	2.57	2.11	1.74	1.33	
18	2.88	2.55	2.10	1.73	1.33	
19	2.86	2.54	2.09	1.73	1.33	
20	2.84	2.53	2.09	1.72	1.32	
30	2.75	2.46	2.04	1.70	1.31	
40	2.70	2.42	2.02	1.68	1.30	
60	2.66	2.39	2.00	1.67	1.30	
120	2.62	2.36	1.98	1.66	1.29	
	2.58	2.33	1.96	1.645	1.28	

α = Level of significance or the % probability that the difference is due to chance alone.

t_c = Critical t-value

Since 52 is outside the 95% confidence level for the experimental mean, the difference between them is significant (suggesting a poor method capability for accuracy). Multiple laboratory test data can provide additional information on method accuracy if the technique is selected for collaborative testing.

3.5 Data Evaluation of Reference Materials Over Time

In conjunction with accuracy data obtained on reference materials, one should establish a control chart. The chart is designed to monitor the consistent performance of a method over time. In such a study a series of measurements are performed on a single sample according to a planned schedule. The frequency of testing (daily, weekly) will be dictated by the particular test method protocol under investigation. Such measurements are plotted on a chart (Fig. 3-1) and provide evidence of repeatability (precision). The chart also indicates whether systematic errors are possibly being introduced by monitoring the consistency of values over time (see Section 4).

The standard deviation of the reference sample is used to define the control limits as follows: two times S (2S) will be the warning limit and three times S (3S) will be the action limit, for deviations outside the 99.7% and 95% confidence limits, respectively. Responses outside the warning limit denote a potential problem in the monitoring system whereas results outside the action limit signal the need for corrective action. The control chart should serve as a tool for checking measurement repeatability for a given method and serve as an index for performance of the laboratory using the test system. When the known response value for the reference material is also plotted, the accuracy of the method may be illustrated by comparison with the experimental values. If the known value is within 2S of the experimental mean throughout the testing period, the data suggest that the method has a good capability for accuracy.

4. SYSTEMATIC ERROR

4.1 Definitions

Systematic error is any non-random deviation in the results of a procedure that influences the accuracy of a method. Inherent systematic or abnormally distributed errors arise from a variety of sources and should be minimized where possible to more closely approximate true values.

An expression of systematic error (bias) can be considered as a part of a method's capability for accuracy. As stated before (Section 3, Method Accuracy), in order for a single laboratory test to estimate the method's capability for (minimizing) systematic error, there must be a reference material and a "true response" must be known for this reference sample. When this situation exists, various concentrations of the reference material can be tested and the results compared (precision, accuracy, etc.). Within a given sample matrix, decreasing analyte concentrations may result in a corresponding loss of accuracy. However, when the technique being evaluated is a toxicity test, there is usually no true value and no true response (except a comparison of results with results from another toxicity test which has already been single laboratory and collaboratively tested). The comparison of results with results of a reference method is not addressed in this guidance document.

Some general sources of systematic error in any analytical method are: 1) insufficient selectivity, the inability to resolve the component of interest from contaminating components which also elicit a response, which gives rise to falsely high results and 2) interferences by components which do not propagate a signal by themselves but which inhibit or enhance the test reading, and, 3) time dependent errors (drift).

These sources of systematic error may be caused by:

1) impure or unstable reagents,
2) instrumental drift, varying zero point, and other equipment instabilities,
3) subjectivity in the reading of endpoints,
4) inadequate determination of random error distribution (poor precision),
5) when applicable, insufficient standardization of the test organism (e.g., unspecified age and feeding schedules).

The sources of systematic error may be reduced by strict adherence to the test protocol. Note: The purpose of a single laboratory test is to evaluate a method (a written protocol), not to evaluate the laboratory conducting the single laboratory test. A major responsibility of the test laboratory is to strictly follow the method protocol. If the method protocol does not provide the following type of information, the sponsor should be notified so that the protocol can be revised before testing is initiated.

1) Label all reagents with expiration date and storage conditions. Do not use reagents which have expired or were stored under conditions other than noted on the label or in the protocol. Use reagents of known purity or grade (as specified in protocol).

2) Maintain all instruments according to protocol instructions and to their instruction manuals. A rigid maintenance program will reduce instrument failure; if applicable, measure blank values and standard values often throughout the analysis (as indicated in the protocol).

3) Conduct preliminary trials before the actual single laboratory test begins.

4) If confusion exists concerning any portion of the test protocol, contact the sponsor immediately.

4.2 Identification of Systematic Errors

The systematic error may be identified simultaneously with the tests establishing the accuracy and precision of the test method. As stated before, all analyses should be planned in advance.

As previously discussed, a true response value for a given reference material may not be available for many biological tests. In these cases no comparisions between experimental and accepted values can be made, and the methods' single laboratory capability for accuracy (and systematic error) cannot be identified.

Although the error cannot be identified in these cases, its presence may be discovered by plotting the data used to establish a known value on a modified control chart, (see Section 3.1).

For cases where a true value is available for a given reference material (i.e., bioanalytical methods) the systematic error will be identified as follows:

Prepare a minimum of 3 different concentrations of the reference material and perform 10 successive, independent analyses on each concentration. Each analysis must be performed with the number of replicates and controls required by the test protocol to produce a valid response.

Analyze the data gathered during the above tests in the same manner as that previously described in the accuracy section (see Section 3) and equation 3-2.

$$t_{(\alpha = 5,)\ (k = 9)} = \frac{|\bar{X} - R|\sqrt{n}}{S_{\bar{X}}} \qquad \text{Eq. 4-1}$$

The sign of the difference ($\bar{X}$ - R) will signify the direction of the measured error.

If the two-sided critical t value (t_c) is calculated to be greater than that noted for 9 degrees of freedom (10-1) and α = 5%, (greater than 2.26, see Table 3-1), then the difference between the experimental and reference value may be considered to be significant (there is a 5% probability that the differences noted are a result of chance alone) and there is evidence of systematic error. If the t_c value is less than 2.26 then the difference is not considered to be significant and the systematic error is probably small enough to be disregarded.

A constant + or - difference of $\bar{X}$ - R observed at each concentration tested will identify the presence of systematic error. If this is observed, the evaluating laboratory should look for sources of systematic errors in the method by reviewing such items as were listed in Section 4.1.

During the performance of the test method over a period of time the emergence of a systematic error may be observed using a control chart as illustrated in Figure 3-1. If a constant + or - trend in results is observed a systematic error may be developing. These observations should be included with a control chart for each method when testing for accuracy, systematic error or establishing a "true" value.

4.3 Reporting Systematic Error

Tabulate the results of the systematic error tests noting the difference between the experimental and reference value ($\bar{X}$ - R), the calculated critical t value (t_c), at the 95% level of significance, the 95% confidence interval about each experimental result (see section 3.4), and the significance of the observed difference for each reference sample concentration tested.

5. METHOD SENSITIVITY

5.1 Definitions and Procedures

For purposes of a single laboratory test, a method's sensitivity is defined as the method's capability to detect (or distinguish between) small changes in sample concentration (i.e., concentration of analyte). A chemically characterized reference material should be used as sample material during sensitivity testing.

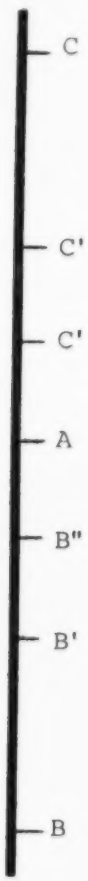


Figure 5-1. Example of different reference material concentrations used in sensitivity testing.

Assume concentration A (Figure 5-1) had been selected as the concentration of sample material used previously in the method precision test. For the sensitivity test, the laboratory would select one concentration greater than (C, Figure 5-1), and one concentration less than (B, Figure 5-1) the concentration used during the precision test. These concentrations should be equally distant from the precision test concentration (A). The laboratory should conduct 10 independent analyses for each new concentration (i.e., 10 separate valid responses acquired by following the method protocol). Assuming that the procedure can distinguish between A and C, and between A and B, then the test laboratory will reduce the concentration interval by one half (to C' and B'). If the method is still capable of distinguishing between A and C', and between A and B', then the test laboratory will again reduce the sample concentration interval by one half (to C" and B"). The 10 independent analyses on these last respective concentrations (C" and B") will complete the sensitivity test even if the method can distinguish between A and C", and between A and B". The sponsor can indicate if additional information on the method's sensitivity is required and if so, direct the test laboratory to continue this process or to repeat the process using a different reference material.

The statistical test used to determine method sensitivity will be the Student's t-Test (see Sections 3 and 4). When the results for the concentrations tested are not significantly different, then the sensitivity limits of the method have been reached. As previously stated in Sections 3 and 4, significance is noted at the 95% level of confidence.

For biological test systems, sensitivity carries an additional importance since different test substances or mixtures of test substances often produce different sensitivities for a given method. The measured sensitivity of a biological system to a test substance or mixture should never be used as a general statement for method sensitivity. The test material used for the sensitivity test must also be stated.

Sensitivity testing can be simultaneously performed with accuracy and systematic error testing to minimize time and effort. The precision section must be completed prior to evaluating sensitivity since precise experimental mean response values are required.

5.2 Reporting Method Sensitivity

To report method sensitivity, the following information will be included:

1. the test material and the concentrations used,
2. individual elicited response values,
3. calculated mean values with standard deviations,
4. results of all t-Tests including 95% confidence intervals,
5. concentrations at which no significant differences (test results) were demonstrated, if sensitivity test was continued to this stage.

6.0 LIMITS OF RELIABLE MEASUREMENT

6.1 Estimating Limits of Reliable Measurement

In determining a method's limits of reliable measurement, the single laboratory test data may simply verify that the method capabilities for sensitivity, precision, and accuracy (if applicable) do not deteriorate at the upper and lower extremes of the detection range. If, in other instances, deterioration in these parameters are noted, then this test data will have demonstrated that the limits of reliable measurement are more narrowly defined than was originally thought.

The single laboratory test is not required to establish an upper and lower detection limit. It is assumed that a sufficient literature data base for the method exists to grossly estimate an upper and lower limit of detection using the particular sample material. The same reference material should be used for these tests as was originally used during the method precision and method sensitivity evaluations.

The evaluating laboratory should initially select two concentrations of the sample material. One of these concentrations will be near the upper extreme of the method's detection range and the other concentration will be near the lower extreme of the detection range. Ten analyses should be conducted on each concentration to provide precision data (expressed as a coefficient of variation). The coefficient of variation will frequently show a dramatic increase at the extreme limits of detection and, therefore, precision data will provide a distinct indication of the limits of reliable measurement. Two additional concentrations, one at each extreme of the estimated response range should then be selected in order to conduct the sensitivity determinations. Thus, two sample concentrations at each extreme of the estimated response range can be compared (in terms of precision and sensitivity) with the previously acquired data. Additional sample concentrations, or additional test substances, can also be selected based on the sponsor's needs. If the method is a bioanalytical technique with an available "true value" for the reference material, accuracy data would also be acquired.

6.2 Example of Reliable Measurement Estimate

The following is a simplistic example given for illustrative purposes only. Under actual test conditions, a limit of reliable measurement may not actually be established. However, even under these conditions, data would still be available to indicate that the technique was (or was not) capable of providing reliable measurements at the extreme concentrations selected.

For this example, three concentrations (or dilutions) are selected at the lower portion of the estimated response range (as per discussion with the sponsor). These concentrations are 1_{μ} g/ml, 100_{μ} g/ml, and 150_{μ} g/ml. Previously acquired precision and sensitivity data have been acquired near the estimated center of the response range (1-100 µg/ml for this particular test material).

The precision data, expressed as a coefficient of variation (CV), from samples with 1 µg/ml are approximately 25 percent greater than the CV of the original precision test, acquired from the samples with concentrations near the estimated center of the response range. Precision data from tests using 100 µg/ml and 150 µg/ml are approximately 10 percent greater than the CV from the original precision test.

In addition, while the method does respond when using the 1 µg/ml sample, the same test response is noted at 100 µg/ml, i.e., the method does not distinguish, in any detectable way, between these two samples. However, a different response is noted at 150 µg/ml indicating, for example, an increased concentration of analyte, a greater inhibition of an enzyme, a greater inhibition of an algal population, etc. Therefore, based on precision and sensitivity concentrations, 100 µg/ml would be the suggested lower limit of reliable measurement when using this particular reference material.

Obviously, this example would not apply to all tests since the possiblity of detecting a mutagen may be of value even if the precision and sensitivity capabilities are not impressive. Knowledge of the method's intended use is essential to all single laboratory tests.

6.3 Reporting the Limits of Reliable Measurement

1) An initial estimate of the upper and lower concentrations of the response range were based on the available literature data base. These sources of data should be listed.

2) The chemical composition of all sample material used should be included as well as a compilation of data from the various precision and sensitivity tests. Accuracy data at the upper and lower extremes in concentration would also be presented, if available. The coefficient of variation should be included for each concentration tested. The method's capability for sensitivity, as indicated by the test response at the concentration extremes, should emphasize the concentrations where no change in response was first observed in spite of the change in sample concentration.

3) The intended purpose of the method should be presented again as well as a statement of what the single laboratory test has concluded about the method's limits of reliable measurement.

7. DATA REQUIRED FOR A COLLABORATIVE TEST DECISION

During the collaborative (or interlaboratory) study, different laboratories analyze aliquots of the same test material in order to validate the performance of a given set of analytical procedures (a written protocol). Data from these collaborative tests are used to determine the reproducibility (overall between-laboratory variability) that can be expected when the method is used by qualified laboratories. The collaborative test is ultimately necessary before a method can be reliably used (e.g., biological tests used to implement RCRA requirements as well as most monitoring techniques as per EPA mandatory quality assurance requirements). However, collaborative testing is expensive and time consuming. It is important that single laboratory test data be available as a basis for selecting which methods will be collaboratively tested. The single laboratory test must characterize a method as to its capability for precision, sensitivity, limits of reliable measurement and ruggedness. Accuracy and systematic error must be evaluated whenever possible (that is, when true values are available).

Therefore, depending on the procedure being evaluated, single laboratory testing should provide the following information.

1) The precision (or repeatability) of the test method for evaluating a known material, expressed as the standard deviation from the mean experimental value.

2) The accuracy of the method for analyzing a known material expressed as percent difference of the mean from the reference value. If no reference value is available, an average of 20 successive independent measurements on aliquots of the same reference material must be performed to establish a known value.

3) The sensitivity of the test method. The sensitivity will be expressed as the method's capability to distinguish between small concentration changes of the same test material.

4) The limits of reliable measurement using reference sample material. The testing laboratory will estimate the upper and lower concentrations of the method's detection limit based on the available literature. These literature sources of data must be listed (non-published data can also be used when making these estimates). Precision, sensitivity, and (if applicable) accuracy data will then be determined for the selected reference samples using concentrations from the upper and lower extremes of the detection range.

5) Procedural variables which must be carefully controlled (and provides a revised protocol).

6) When the "true response" is available for a reference material, the single laboratory test will provide an estimate of the method's capability for minimizing systematic error.

The single laboratory test cannot determine a method's scope of applicability, i.e., the array of sample types or environmental situations for which the method would be able to provide useful data. However, the single laboratory test should include some test samples having a concentration range in the same sample matrix that would be encountered if the method was being routinely used for its intended purpose.

The method being tested can be improved during the single laboratory evaluation process (especially during the ruggedness testing phase) but such improvements must be incorporated into a revised protocol. The evaluating laboratory must strictly follow the written procedure throughout the single laboratory test if the resulting data are to be used as a basis for proceeding to collaborative testing.

APPENDIX I. OUTLIERS - REJECTION OF SUSPECTED DATA

A valid estimation of a mean and its respective standard deviation must include the elimination of outliers. Outliers are observed results which lie outside of the statistical frequency distribution in a series of analytical values, i.e., a value so far removed from the other values that the laboratory suspects that it may be in error. Obviously, if the experimenter has evidence that a result was caused by an oversight, mistake, or malfunction, the observation in question should be rejected. However, one should not indiscriminately reject observations which appear to be more widely spaced than one would like. For example, if large deviations occur in a number of samples, one should suspect the presence of additional factors such as systematic error (see Section 4).

The outlier, by definition, is a result which is manifestly inconsistent with other data in a set, and one which may or may not be in error. Its occurrence is assumed to be caused by a gross error, in which no specific knowledge as to the nature of the mistake is evident. One should expect to find that when a variate is normally distributed, 95 percent of all observations fall within 1.96 true standard deviations, see Table A-4, of the mean (Neville and Kennedy, 1964). That is to say, the chance of a value falling outside of that range is 1 in 20 (for a 99 percent confidence level, the chance of falling outside the range is 1 in 100, and so on).

It is then reasonable to reject observations which differ by more than a specified confidence interval. The confidence interval chosen depends on the size of the sample. The 95 percent confidence level is generally the lowest acceptable confidence level.

A number of criteria for the rejection of outliers have been proposed. Two example methods are listed below.

Example:

The first example is based on results from a biological assay conducted at different dose levels and on ten different days, and is shown in Table A-1. A quick glance at Table A-1 suggests that the value 16.5 obtained at 4 mg/ml is unusually far removed from the other data. A statistical test for the elimination of suspected outliers is performed according to the procedure described by Dixon and Massey (1951). According to this scheme, a set of n observations taken from a normally distributed population, in this case 30 for each dose, are ranked in order from the largest to the smallest. If the observation in question was the smallest value, the order of ranking would be reversed (i.e. smallest to largest).

16.5	8.9	8.9	8.8	8.8	8.8	8.7	8.6	8.6	8.6	8.6	8.6	8.5	8.5	8.5
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
8.5	8.5	8.5	8.5	8.4	8.4	8.3	8.3	8.3	8.3	8.2	8.2	8.2	8.1	8.0
X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}

Hypothetical results of a biological monitoring assay conducted on ten different days to estimate the precision of the assay illustrates the example (see Tables A-1).

TABLE A-1. RESULTS OF A BIOLOGICAL MONITORING ASSAY

Dose Level or Conc. (mg/ml)	Days A			B			C			D			E		
1	1.0	1.7	2.2	1.5	2.8	1.8	1.2	0.9	1.3	1.4	1.2	1.8	1.3	1.5	1.7
2	4.6	5.2	4.8	5.2	5.6	5.2	4.7	5.4	5.3	5.1	5.0	5.0	4.6	4.8	4.7
4	8.5	8.2	8.6	8.3	8.4	16.5	8.9	8.2	8.5	8.3	8.4	8.6	8.5	8.5	8.7
6	10.8	13.8	11.5	12.8	10.2	11.3	13.0	12.6	11.7	11.4	10.9	11.7	12.0	11.8	11.7
8	14.6	14.8	16.9	14.9	14.9	15.8	16.5	16.2	17.3	14.9	15.5	15.4	14.9	15.6	16.2

Dose Level or Conc (mg/ml)	Mean ($\bar{X}$)	St. Dev. (S)	Coef. Var (V,%)
1	1.5	0.40	26.7
2	5.0	0.31	6.2
4	8.8[a]	1.48[a]	16.8[a]
6	11.8	0.71	6.0
8	15.4	0.92	6.0

[a] Values prior to rejection of outlier. Corrected values are 8.5, 0.23, and 2.7% respectively.

TABLE A-1 continued

F			G			H			I			J		
1.0	1.4	1.6	0.9	1.3	1.2	1.6	1.7	1.7	1.8	1.5	1.3	2.0	1.7	1.6
4.5	5.2	4.9	4.9	5.0	5.2	5.4	5.5	4.9	4.6	4.9	5.0	5.0	4.5	4.5
8.5	8.6	8.3	8.8	8.5	8.8	8.1	8.3	8.2	8.0	8.5	8.6	8.8	8.9	8.6
11.7	11.8	11.5	12.0	10.9	10.9	12.0	12.1	12.2	11.5	11.6	11.9	12.1	12.5	11.9
16.0	14.9	15.2	15.6	12.0	14.9	15.5	16.0	15.1	15.1	16.0	16.0	15.3	15.4	15.5

The appropriate formula for n=30, is defined as $r_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$, (see Table A-2)

and in this example: $r_{22} = \frac{8.9-16.5}{8.2-16.5} = 0.916$

The computed experimental r value is compared to the critical r value obtained from Table A-2. In this case, the calculated r_{22} exceeds the tabulated r_{22} of 0.457 at a 1 percent level of significance. Therefore, the suspected deviant value can be rejected with 99% confidence and the data set for the dose of 4 mg/ml is re-evaluated excluding the outlier 16.5 and a revised X and S are obtained. The corrected standard deviation for each concentration, outliers eliminated, will be taken as an estimate of method precision. The relative precision at each concentration will be calculated as the coefficient of variation (CV).

$$CV = \frac{S}{\bar{X}} (100) \qquad \text{(Eq. 2-2, Section 2)}$$

Table A-2. * CRITERIA FOR TESTING FOR EXTREME OBSERVATION (no past data)

Statistic	Sample Size n	Critical Values α = 5 percent	α = 1 percent
$r_{10} = \frac{x_2 - x_1}{x_n - x_1}$	3	0.941	0.988
	4	0.765	0.889
	5	0.642	0.780
	6	0.560	0.698
	7	0.507	0.637
$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$	8	0.554	0.683
	9	0.512	0.635
	10	0.477	0.597
$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1}$	11	0.576	0.679
	12	0.546	0.642
	13	0.521	0.615
$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}$	14	0.546	0.641
	15	0.525	0.616
	16	0.507	0.595
	17	0.490	0.577
	18	0.475	0.561
	19	0.462	0.547
	20	0.450	0.535
	21	0.440	0.524
	22	0.430	0.514
	23	0.421	0.505
	24	0.413	0.497
	25	0.406	0.489
	26	0.399	0.486
	27	0.393	0.475
	28	0.387	0.469
	29	0.381	0.463
	30	0.376	0.457

*Adapted from W. J. Dixon and F. J. Massey, Introduction to Statistical Analysis (McGraw-Hill Book Company, Inc., New York, 1969).

Example: Rejection on the Basis of Estimated Variance

This criterion is called Nair's method. It states that the maximum deviation from the sample mean, $\bar{X}$, that can be expected for single values in n sample is related to the estimated variance of the population.

For instance, let X_m be the largest or smallest value of X that can be expected in a sample of size n at a given probability level. The probability levels are 10, 5, 1 and 0.1 percent. Let S_e be the estimated standard deviation, the estimate based on data with k degrees of freedom. Table A-3 gives values of $\frac{X_m - \bar{X}}{S_e}$ which are not normally expected to be exceeded in samples of size n between 3 and 9. If a deviate larger than the tabulated value is observed, then the observation may be rejected as being significantly different from the rest of the sample. The level of significance represents the risk of error. (e.g., a 5 percent level of significance means that there is a 5 percent probability that we are rejecting a value which really does belong to the sample population).

For instance: 25 observations on the weights of material from a packing machine have given an estimate of standard deviation as 0.40 lb. Assume that weights, in pounds, taken by a customer on five packages from his consignments are 100.4, 100.2, 100.5, 100.2, 99.2. The question is whether the value of $X_m = 99.2$ can be discarded.

Since $\bar{X} = 100.1$ and $S_e = 0.40$

$$\frac{X_m - \bar{X}}{S_e} = \frac{0.9}{0.40} = 2.25$$

The number 2.25 can then be entered into Table 3 for n=5 and the number of degrees of freedom k = 25-1 = 24. The extreme deviates in Table A-3 are 2.23 at the 5 percent significance level and 2.8 at the 1 percent level. Because the value 2.25 is between the values 2.8 and 2.23, one can reject 99.2 with a 5 percent risk of making a wrong decision.

Having eliminated the extreme deviate, the sample mean is 100.32 lb. and the standard deviation is 0.15 ($\sqrt{0.0225}$). This is based on the number of degrees of freedom k = 4-1 = 3.

TABLE A-3. VALUES OF EXTREME DEVIATE $\frac{X_m - \bar{X}}{S_e}$ NOT REJECTED AS AN OUTLIER

k \ n	3	4	5	6	7	8	9
			α = 5 percent level				
10	2.02	2.29	2.49	2.63	2.75	2.85	2.93
11	1.99	2.26	2.44	2.58	2.70	2.79	2.87
12	1.97	2.22	2.40	2.54	2.65	2.75	2.83
13	1.95	2.20	2.38	2.51	2.62	2.71	2.79
14	1.93	2.18	2.35	2.48	2.59	2.68	2.76
15	1.92	2.16	2.33	2.46	2.56	2.65	2.73
16	1.90	2.14	2.31	2.44	2.54	2.63	2.70
17	1.89	2.13	2.30	2.42	2.52	2.61	2.68
18	1.88	2.12	2.28	2.41	2.51	2.59	2.66
19	1.87	2.11	2.27	2.39	2.49	2.58	2.65
20	1.87	2.10	2.26	2.38	2.48	2.56	2.63
24	1.84	2.07	2.23	2.35	2.44	2.52	2.59
30	1.82	2.04	2.20	2.31	2.40	2.48	2.55
40	1.80	2.02	2.17	2.28	2.37	2.44	2.51
60	1.78	1.99	2.14	2.25	2.33	2.41	2.47
120	1.76	1.97	2.11	2.21	2.30	2.37	2.43
	1.74	1.94	2.08	2.18	2.27	2.33	2.39
			α = 1 percent level				
10	2.76	3.05	3.25	3.39	3.50	3.59	3.67
11	2.71	3.00	3.19	3.33	3.44	3.53	3.61
12	2.67	2.95	3.14	3.28	3.39	3.48	3.55
13	2.63	2.91	3.10	3.24	3.34	3.43	3.51
14	2.60	2.87	3.06	3.20	3.30	3.39	3.47
15	2.57	2.84	3.02	3.16	3.27	3.35	3.43
16	2.55	2.81	3.00	3.13	3.24	3.32	3.39
17	2.52	2.79	2.97	3.10	3.21	3.29	3.36
18	2.50	2.77	2.95	3.08	3.18	3.27	3.34
19	2.49	2.75	2.92	3.06	3.16	3.24	3.31
20	2.47	2.73	2.91	3.04	3.14	3.22	3.29
24	2.43	2.68	2.85	2.97	3.07	3.15	3.22
30	2.38	2.62	2.79	2.91	3.01	3.08	3.15
40	2.34	2.57	2.73	2.85	2.94	3.02	3.08
60	2.30	2.52	2.68	2.79	2.88	2.95	3.01
120	2.25	2.48	2.62	2.73	2.82	2.89	2.95
	2.22	2.43	2.57	2.68	2.76	2.83	2.88

An outlier is rejected if its value exceeds the tabulated value at the requisite level of significance.

X_m = greatest or smallest value that can be expected in a sample of size n at the given level of significance.

S_e = estimate of standard deviation from a sample with k degrees of freedom.

TABLE A-3. (continued)

k \ n	α = 0.1 percent level						
	3	4	5	6	7	8	9
10	3.54	3.84	4.04	4.17	4.28	4.35	4.40
11	3.49	3.80	3.99	4.12	4.23	4.30	4.36
12	3.45	3.75	3.94	4.07	4.19	4.26	4.31
13	3.41	3.71	3.90	4.03	4.14	4.22	4.28
14	3.38	3.67	3.86	4.00	4.10	4.18	4.24
15	3.35	3.64	3.83	3.96	4.06	4.15	4.21
16	3.32	3.51	3.80	3.93	4.03	4.12	4.18
17	3.29	3.58	3.77	3.90	4.00	4.09	4.15
18	3.27	3.55	3.74	3.88	3.98	4.06	4.12
19	3.25	3.53	3.72	3.85	3.95	4.03	4.10
20	3.23	3.51	3.70	3.83	3.93	4.01	4.08
24	3.16	3.44	3.62	3.75	3.85	3.93	4.00
30	3.08	3.36	3.53	3.66	3.76	3.84	3.90
40	3.01	3.27	3.44	3.57	3.66	3.74	3.81
60	2.93	3.19	3.35	3.47	3.56	3.64	3.70
120	2.85	3.10	3.26	3.37	3.46	3.53	3.59
	2.78	3.01	3.17	3.28	3.36	3.43	3.48

TABLE A-4. *USE OF NORMAL DISTRIBUTION
Values of S for A Specified Percentage of Results
To Lie Within the Range $\overline{X} \pm S$

Level of Significance (percentage of results outside the range)	S	Confidence (percentage of results within the range)
10	1.645	90
5	1.960	95
2	2.326	98
1	2.576	99
0.1	3.291	99.9
0.01	3.891	99.99

X = value of independent variable

*adapted from Neville and Kennedy, 1964

APPENDIX II ACCEPTANCE CRITERIA FOR ADEQUATE METHOD PRECISION.

Methods that are intended for use in the biological monitoring of hazardous materials should have a capability for precision such that one can clearly differentiate between normal, acceptable levels and unacceptable levels.

For instance, a hazardous substance will elicit a mean response ($\bar{X}_c$) value (i.e., toxicity measurement) when assayed using a given test method. If for regulatory reasons, the upper limit for responses from the same type of materials is fixed at X_u, then one must be able to discriminate reliably between those two values. In this case, the required discriminatory ability is given by the expression:

$$D_R = \left[\frac{X_u - \bar{X}_c}{\bar{X}_c}\right] x\ 100 \qquad \text{Eq. A-II-1}$$

From Eq. 1-1 estimate of standard deviation, it is apparent that the uncertainty of an analytical value can be minimized by decreasing the standard deviation (S) or by increasing the number of samples (n). The use of the concept of required discrimination can be illustrated with the following example. Consider, for instance, that in a pilot study a consensus value or response $\bar{X}_c = 100$ has been obtained for some particular substance(s), and that the upper regulatory limit for this substance(s) has been fixed at $X_u = 110$; then the required discriminatory ability is

$$D_R = \left[\frac{110-100}{100}\right] 100 = 10\%$$

Or put in another way, the precision of the test must allow the discrimination of a difference of 10% between the measured value and the regulatory limit value.

During the analysis, it is important that the precision of the method (P_M) be larger than the required precision (P_R), that is,

$$P_M > P_R$$

For new methods for which the variablity has not been established, or no sufficient data is available for an accurate determination of the population variance, the required discriminatory ability of the method can be described by the expression

$$D_M = \left[\frac{(S)\ (t_{(\alpha,\ k)})}{(\bar{X}_c)\ \sqrt{n_e}}\right] x\ 100 \qquad \text{Eq. A-II-2}$$

where: S = estimated standard deviation of the mean for the pilot study with n replicates.

α = critical t value from a two-sided Student's t distribution where:

k = $n-1$, where n = number of samples assayed for S.

α = level of significance (see Table 3-1, Percentile Values for Student's t distribution).

n_e = the number of samples that will be assayed for monitoring purposes.

$\bar{X}_c$ = mean response value of pilot study.

Since the discriminating ability (D_M) of the method is proportional to the estimated standard deviation (S) and since the precision of the method (P_M) is inversely proportional to S, D_M must be less than D_R for a method to be suitable for testing. The following example illustrates this point.

Example:

In a pilot study conducted with 10 replicates we find a standard deviation S=10.0; we are planning to run routinely no more than n_e=2 assays and we want to use a 5% level of significance. Then, k=10-1 or 9, and using Table 3-1, $t_{(95\%,\ 9)} = 2.26$ (two sided).

Replacing in Eq. A-II-2,

$$D_M = \frac{10.0 \times 2.26}{100 \times 2} \times 100 = 16\%$$

Since $D_R = 10\%$, then: $D_M > D_R$

This would indicate that in the planned protocol conditions, the method cannot be used for monitoring purposes, since it is not precise enough to meet the required criterion for discriminatory ability. However, the precision of the measurement may be improved (a) by decreasing the variability of the results, S, which can be studied by performing a ruggedness test (Section 1), or (b) by increasing the sample size (n_e) during the routine application. If, for instance, n_e increased from 2 to 6 replicates the precision will be improved from 16 to 9%. However, by duplicating the number of assays to $n_e = 4$, and simultaneously improving the variability of the results by one half (S=5); then the discriminatory ability will decrease to less than 6%.

In summary, in order for a biological monitoring method to meet the acceptance criteria for method precision, the discriminatory ability of the method (D_M) must be less than the required discriminatory ability (D_R).

GLOSSARY OF TERMS

1. ACCURACY - Agreement of a measurement with the "true" (accepted) value of a known material.

2. COEFFICIENT OF VARIATION - A measure of the dispersion of a collection of data around a mean value; usually expressed on a percentage basis i.e., in relative rather than absolute terms.

$$CV = \frac{S}{\bar{X}} \times 100$$

3. COLLABORATIVE TEST - Different laboratories analyze aliquots of the same sample material in order to validate the performance of a given set of analytical (or bioassay) procedures. Data from these collaborative tests are used to determine the reproducibility (overall between-laboratory variability) that can be expected when the method is used by qualified laboratories.

4. COMPLETENESS - A statement of the amount of valid data needed to meet the objectives of the project.

5. CONSENSUS VALUE - A mean value acquired by collecting data on the same test substance and test system from several different laboratories.

6. DATA VALIDATION - The process whereby data are filtered and accepted or rejected based on a set of criteria.

7. HISTORICAL DATA - Test results that are compiled over a period of time from which a mean value and standard deviation can be calculated.

8. LIMITS OF RELIABLE MEASUREMENTS - The upper and lower limits of reliable measurements of the test system for a given method. Lower response limit in biological monitoring methods is the concentration of test substance which elicits the same response from the test system as the negative control. It can also be defined as the concentration (or concentration range) at which the coefficient of variation begins to increase rapidly (imprecision), delineating a lower limit of usefulness for the method. The upper limits of reliable measurement can be regarded as the concentration (or range) at which the test system's response is consistent (and precise) with a maximum response. At concentrations beyond that point, the test response will not be different.

9. OUTLIER - Values that do not follow the developing trend of other data values in a group, or those values that are manifestly inconsistent with other data in a set the probability that the value is not representative of the overall population of values).

10. PRECISION - Is a measure of the reproducibility and/or repeatibility of multiple determinations using the same sample material (how close the measured values are to each other).

11. RANDOM ERRORS - Random errors occur in no set pattern. They can be positive or negative in a given group of results, or may fluctuate in a random manner. Random errors can be caused by incidental variations such as a lapse in technique or equipment performance.

12. REFERENCE MATERIALS - Test samples of known, certified chemical composition or which consistently evoke a known response (true response) from a given test system. These materials must be readily available to test laboratories.

13. RELIABILITY - Of measured values requires that both the accuracy and the precision be within acceptable limits.

14. REPEATIBILITY - Is the measure of within-laboratory random error and is expressed in terms of variability in values of replicate determinations on same day and between day trials.

15. REPRODUCIBILITY - Is the dispersion around the mean obtained when a sample is analyzed by the same method by several laboratories and is considered to be the measure of overall between-laboratory variability, including within-laboratory variability.

16. RUGGEDNESS - How well the results of testing in a method hold true when procedural variations likely to affect the results are introduced during the testing.

17. SENSITIVITY - The method's capability for detecting small changes in pollutant (analyte) concentration.

18. SINGLE LABORATORY TEST - A single laboratory test establishes the data quality that can be achieved within a single laboratory (using a specified procedure), it provides a basis for deciding whether or not a given method merits collaborative testing, and it more clearly defines a method's potential for inclusion as part of an operational monitoring network.

19. SPIKED SAMPLES - Sample materials to which a known amount of analyte has been added. The amount of material added should not be excessive in relation to the amount present.

20. SYNTHETIC SAMPLES - Laboratory prepared products with known amount of analyte by actual preparation (followed by subsequent confirming analysis).

21. SYSTEMATIC ERRORS - Are produced by variations which influence results consistently in one direction or the other (i.e., yielding all positive or all negative errors).

NOTATION

$\bar{X}$	Estimate of population mean or sample mean
X_i	Value of an independent variable
CV	Coefficient of variation
S	Estimate of standard deviation
n	Sample size
k	Number of degrees of freedom
α	Level of significance, or the percent probability of the observed result being due to chance alone.

REFERENCES

ACS Subcommittee report (1982). "Improving analytical chemical data used for public purposes". Chem. Eng. News, 60 (23):44-48.

Alvarez, R., S. D. Rasberry and G. A. Uriano (1982). "NBS standard reference materials: Update 1982". Anal. Chem., 54:1226A-1244.

ANSI/ASTM (1978). "Standard recommended practice for developing precision data on ASTM methods for analysis and testing of industrial chemicals". ANSI/ASTM E180-78, 690-714.

ASTM (1980) "Standard Practice for conducting an interlaboratory test program to determine the precision of test method". ASTM E-691, 959-992.

Branson, D. R., D. N. Armentrout, W. M. Parker, C. Van Hall, and L. I. Bone (1981). "Effluent monitoring step by step". Environ. Sci. Technol., 15:513-518.

Capar, S. G. and C. A. Subjoc (1982). "Defining a lowest level of reliable measurement for lead in foods". AOAC, 65: 1025-1029.

Devesis, E., J. P. Boehler, T. E. Givand, and J. C. Sheridan (1982). "Submitting HPLC methods to the compendia and regulatory agencies". Pharm. Technol., 6(9):120-137.

Dixon, W. J. and F. J. Massey (1951). "Efficiency and various statistics" in Introduction to Statistical Analysis. McGraw-Hill Book Company, Inc., 127-149.

Eisenhart C. (1968). "Expression of the uncertainties of final results". Science, 160:1201-1204.

Hertz, H. S., L. R. Hilert, W. E. May, S. A. Wise, J. M. Brown, S. N. Chesler and F. R. Guenther (1979). "Interlaboratory comparison of environmental analyses associated with increased energy production". ASTM/STP-686, 291-301.

Horwitz, W. (1982a). "Evaluation of analytical methods used for regulation". AOAC, 65: 525-530.

Horwitz, W. (1982b). "Evaluation of analytical methods used for regulation of food and drugs". Anal. Chem., 54:67A-76A.

Iyengar, V. (1982). "Presampling factors in the elemental composition of biological systems". Anal. Chem., 54:554A-559A.

Kratochvil, B. and J. K. Taylor (1981). "Sampling for chemical analysis". Anal. Chem., 53:924A-938A.

Litchfield, J. T. and F. Wilcoxon (1949). "A simplified method of evaluating dose-effect experiments". J. Pharmacol. Exp. Ther., 96:99-116.

MacDougall, D., and Crummett, W. B. (1980). "Guidelines for data acquisition and data quality evaluation in environmental chemistry". Anal. Chem., 52:2242-2249.

Miller, S. (1982). "Quality assurance, analytical methods, and hazardous wastes". Environ. Sci. Technol., 16: 332A-336A.

Neville, Adam M. and John B. Kennedy (1964). "Rejection of outliers" in Basic Statistical Methods for Engineers and Scientists, International Textbook Company, 125-128.

Rice, J. K. (1980). "Analytical issues in compliance monitoring". Environ. Sci. Technol., 14:1455-1457.

Snedecor, George W. and William Cochran (1968). "The comparison of two samples" in Statistical Methods, The Iowa State University Press, 91-119.

Taylor, J. K. (1981). "Quality assurance of chemical measurements". Anal. Chem., 53:1588A-1596A.

Vanderwielen, A. J. and E. A. Hardwidge (1982). "Guidelines for assay validation". Pharm. Technol., 6(3):66-76.

Youden, W. J. (1969) "The collaborative test" in Precision Measurement and Calibration, H.H.Ku Ed., US Dept. of Commerce, National Bureau of Standards 151-158.

Youden, W. J. and E. H. Steiner (1975). Statistical Manual of the Association of Analytical Chemists, AOAC, 88pp.